C1s1680-1

日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００１年　３月　７日

出願番号
Application Number: 特願２００１－０６３９６８

[ST.10/C]: [ＪＰ２００１－０６３９６８]

出願人
Applicant(s): 鈴木　昌和
株式会社東芝

２００２年　１月１１日

特許庁長官
Commissioner,
Japan Patent Office 及川耕造

出証番号　出証特２００１－３１１５２０７

【書類名】　特許願

【整理番号】　A000100951

【提出日】　平成13年 3月 7日

【あて先】　特許庁長官　殿

【国際特許分類】　G06K 9/00

【発明の名称】　数式認識装置および数式認識方法並びに文字認識装置および文字認識方法

【請求項の数】　17

【発明者】

【住所又は居所】　福岡県福岡市東区箱崎６丁目１０番１号　九州大学内

【氏名】　鈴木　昌和

【発明者】

【住所又は居所】　福岡県福岡市東区箱崎６丁目１０番１号　九州大学内

【氏名】　江藤　裕子

【発明者】

【住所又は居所】　東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内

【氏名】　横田　和章

【特許出願人】

【住所又は居所】　福岡県福岡市東区箱崎６丁目１０番１号　九州大学内

【氏名又は名称】　鈴木　昌和

【特許出願人】

【識別番号】　000003078

【氏名又は名称】　株式会社　東芝

【代理人】

【識別番号】　100058479

【弁理士】

【氏名又は名称】　鈴江　武彦

【電話番号】　03-3502-3181

【選任した代理人】

【識別番号】　100084618

【弁理士】

【氏名又は名称】　村松　貞男

【選任した代理人】

【識別番号】　100068814

【弁理士】

【氏名又は名称】　坪井　淳

【選任した代理人】

【識別番号】　100092196

【弁理士】

【氏名又は名称】　橋本　良郎

【選任した代理人】

【識別番号】　100091351

【弁理士】

【氏名又は名称】　河野　哲

【選任した代理人】

【識別番号】　100088683

【弁理士】

【氏名又は名称】　中村　誠

【選任した代理人】

【識別番号】　100070437

【弁理士】

【氏名又は名称】　河井　将次

【手数料の表示】

【予納台帳番号】　011567

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9705037

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　数式認識装置および数式認識方法並びに文字認識装置および文字認識方法

【特許請求の範囲】

【請求項１】　数式を含む文書イメージの文字認識を行う文字認識手段と、

正規表現により特定可能な単語種別毎にそれがテキストと数式に該当する可能性をそれぞれ示す評価値を定義した第１の知識辞書と、

前記第１の知識辞書を参照して、前記文字認識手段によって得られた文字認識結果に含まれる各単語についてテキストおよび数式それぞれに該当する評価値を得る手段と、

形式文法と前記各単語毎に算出されるテキストおよび数式それぞれの評価値とに基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路を探索し、その探索結果に基づいて数式に該当する単語を検出する数式検出手段とを具備することを特徴とする数式認識装置。

【請求項２】　接続可能な単語それぞれのテキスト品詞と数式の関係を前記形式文法として定義した第２の知識辞書をさらに具備し、

前記数式検出手段は、

前記文字認識結果に含まれる各単語についての品詞と前記第２の知識辞書で与えられる形式文法とに従って、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続可能な全ての経路を選定し、それら経路の中で、単語それぞれのテキストまたは数式に関する合計評価値が最大となる最適な経路を探索することを特徴とする請求項１記載の数式認識装置。

【請求項３】　前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を、前後の文字種類別に複数記憶する手段と、

前記数式検出手段で検出された数式内に含まれる前後の文字毎に、正規化サイズとその中心位置の関係を算出し、その算出結果と、前記前後の文字の文字種類の関係に対応するサンプル情報とに基づいて、前記前後の文字間毎に、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造

候補とその評価値から成るリンク候補を得る文字間構造判定手段をさらに具備することを特徴とする請求項１記載の数式認識装置。

【請求項４】　前記数式内に含まれる文字それぞれの文字高さの分布に基づいて予め決められた大域的評価条件を記憶する手段と、

前記大域的評価条件と、前記リンク候補とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれかの文字間構造候補を選択しながら前記数式内の文字同士を矛盾なく接続するための最適な経路を探索し、その探索結果に基づいて前記各文字間の水平位置関係、下付添え字関係、上付添え字関係を認識する手段とをさらに具備することを特徴とする請求項３記載の数式認識装置。

【請求項５】　前記大域的評価条件には、下付添え字領域に含まれる文字の高さと他の各文字の高さとの関係、ベースラインと下付添え字領域に含まれる文字との間の位置関係、水平領域に含まれる文字間の高さのバラツキ、のうちの少なくとも１つが含まれていることを特徴とする請求項４記載の数式認識装置。

【請求項６】　前記数式検出手段で検出された数式をその数式構成要素毎に分解し、各数式構成要素から少なくとも左添え字、アクセント記号、根号、点類を検出して、それを除外する手段をさらに具備し、

前記文字間構造判定手段は、除外した数式構成要素に対して、リンク候補を得ることを特徴とする請求項３記載の数式認識装置。

【請求項７】　数式を含む文書イメージの文字認識を行う文字認識手段と、

前記文字認識手段によって得られた文字認識結果の中から数式領域を検出する手段と、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を、前後の文字種類別に複数記憶する手段と、

前記数式領域内に含まれる前後の文字毎に、正規化サイズとその中心位置の関係を算出し、その算出結果と、前記前後の文字の文字種類の関係に対応するサンプル情報とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係それぞれに該当する可能性を示すリンク候補を得る手段とを具備

することを特徴とする数式認識装置。

【請求項８】　数式を含む文書イメージの文字認識を行う文字認識手段と、

前記文字認識手段によって得られた文字認識結果の中から数式領域を検出する手段と、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を記憶する手段と、

前記数式領域内に含まれる前後の文字毎に正規化サイズとその中心位置の関係を算出し、その算出結果と、前記サンプル情報とに基づいて、前記前後の文字間毎に、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造候補とその評価値から成るリンク候補を得る文字間構造判定手段と、

前記数式領域内に含まれる文字それぞれの文字高さの分布に基づいて予め決められた大域的評価条件を記憶する手段と、

前記大域的評価条件と、前記リンク候補とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれかの文字間構造候補を選択しながら前記数式領域内の文字同士を矛盾なく接続するための最適な経路を探索し、その探索結果に基づいて前記各文字間の水平位置関係、下付添え字関係、上付添え字関係を認識する手段とを具備することを特徴とする数式認識装置。

【請求項９】　数式を含む文書イメージの文字認識を行う文字認識ステップと、

正規表現により特定可能な単語種別毎にそれがテキストと数式に該当する可能性をそれぞれ示す評価値を定義した第１の知識情報を参照して、前記文字認識ステップによって得られた文字認識結果に含まれる各単語についてテキストおよび数式それぞれに該当する評価値を得るステップと、

形式文法と前記各単語毎に算出されるテキストおよび数式それぞれの評価値とに基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路を探索し、その探索結果に基づいて数式に該当する単語を検出する数式検出ステップとを具備することを特徴とする数式認識方法。

【請求項１０】　数式を含む文書イメージの文字認識を行う文字認識ステップと、

前記文字認識ステップによって得られた文字認識結果の中から数式領域を検出するステップと、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を、前後の文字種類別に複数予め用意しておき、前記数式領域内に含まれる前後の文字毎に、正規化サイズとその中心位置の関係を算出し、その算出結果と、前記前後の文字の文字種類の関係に対応するサンプル情報とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係それぞれに該当する可能性を示す文字間構造の評価値から成るリンク候補を得るステップとを具備することを特徴とする数式認識方法。

【請求項１１】　数式を含む文書イメージの文字認識を行う文字認識ステップと、

前記文字認識ステップによって得られた文字認識結果の中から数式領域を検出するステップと、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を予め用意しておき、前記数式領域内に含まれる前後の文字毎に正規化サイズとその中心位置の関係を算出し、その算出結果と、前記サンプル情報とに基づいて、前記前後の文字間毎に、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造候補とその評価値から成るリンク候補を得る文字間構造判定ステップと、

前記数式領域内に含まれる文字それぞれの文字高さの分布に基づいて予め決められた大域的評価条件を用意しておき、前記大域的評価条件と、前記リンク候補とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれかの文字間構造候補を選択しながら前記数式領域内の文字同士を矛盾なく接続するための最適な経路を探索し、その探索結果に基づいて前記各文字間の水平位置関係、下付添え字関係、上付添え字関係を認識するステップとを

具備することを特徴とする数式認識方法。

【請求項１２】　数式を含む文書を読み取り、テキスト領域および数式領域それぞれについての認識処理を行う文字認識装置において、

前記数式を含む文書のイメージデータに対して文字認識を行う文字認識手段と、

正規表現により特定可能な単語種別毎にそれがテキストと数式に該当する可能性をそれぞれ示す評価値を定義した第１の知識辞書と、

前記第１の知識辞書を参照して、前記文字認識手段によって得られた文字認識結果に含まれる各単語についてテキストおよび数式それぞれに該当する評価値を得る手段と、

形式文法と前記各単語毎に算出されるテキストおよび数式それぞれの評価値とに基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路を探索し、その探索結果に基づいて前記数式領域と前記テキスト領域を検出する手段とを具備することを特徴とする文字認識装置。

【請求項１３】　数式を含む文書を読み取り、テキスト領域および数式領域それぞれについての認識処理を行う文字認識装置において、

前記数式を含む文書のイメージデータに対して文字認識を行う文字認識手段と、

前記文字認識手段によって得られた文字認識結果の中から数式領域を検出する手段と、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を、前後の文字種類別に複数記憶する手段と、

前記数式領域内に含まれる前後の文字毎に、正規化サイズとその中心位置の関係を算出し、その算出結果と、前記前後の文字の文字種類の関係に対応するサンプル情報とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係それぞれに該当する可能性を示すリンク候補を得て、前記数式領域内の数式構造を認識する手段とを具備することを特徴とする文字認識装置。

【請求項１４】　数式を含む文書を読み取り、テキスト領域および数式領域

それぞれについての認識処理を行う文字認識装置において、

前記数式を含む文書のイメージデータに対して文字認識を行う文字認識手段と、

前記文字認識手段によって得られた文字認識結果の中から数式領域を検出する手段と、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を記憶する手段と、

前記数式領域内に含まれる前後の文字毎に正規化サイズとその中心位置の関係を算出し、その算出結果と、前記サンプル情報とに基づいて、前記前後の文字間毎に、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造候補とその評価値から成るリンク候補を得る文字間構造判定手段と、

前記数式領域内に含まれる文字それぞれの文字高さの分布に基づいて予め決められた大域的評価条件を記憶する手段と、

前記大域的評価条件と、前記リンク候補とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれかの文字間構造候補を選択しながら前記数式領域内の文字同士を矛盾なく接続するための最適な経路を探索し、その探索結果に基づいて前記各文字間の水平位置関係、下付添え字関係、上付添え字関係を認識する手段とを具備することを特徴とする文字認識装置。

【請求項１５】　数式を含む文書を読み取り、テキスト領域および数式領域それぞれについての認識処理を行う文字認識方法において、

前記数式を含む文書のイメージデータに対して文字認識を行う文字認識ステップと、

正規表現により特定可能な単語種別毎にそれがテキストと数式に該当する可能性をそれぞれ示す評価値を定義した第１の知識辞書を参照して、前記文字認識ステップによって得られた文字認識結果に含まれる各単語についてテキストおよび数式それぞれに該当する評価値を得るステップと、

形式文法と前記各単語毎に得られたテキストおよび数式それぞれの評価値とに

基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路を探索し、その探索結果に基づいて前記数式領域と前記テキスト領域を検出するステップとを具備することを特徴とする文字認識方法。

【請求項１６】　数式を含む文書を読み取り、テキスト領域および数式領域それぞれについての認識処理を行う文字認識方法において、

前記数式を含む文書のイメージデータに対して文字認識を行う文字認識ステップと、

前記文字認識ステップによって得られた文字認識結果の中から数式領域を検出するステップと、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を、前後の文字種類別に複数用意しておき、前記数式領域内に含まれる前後の文字毎に、正規化サイズとその中心位置の関係を算出し、その算出結果と、前記前後の文字の文字種類の関係に対応するサンプル情報とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係それぞれに該当する可能性を示すリンク候補を得て、前記数式領域内の数式構造を認識するステップとを具備することを特徴とする文字認識方法。

【請求項１７】　数式を含む文書を読み取り、テキスト領域および数式領域それぞれについての認識処理を行う文字認識方法において、

前記数式を含む文書のイメージデータに対して文字認識を行う文字認識ステップと、

前記文字認識ステップによって得られた文字認識結果の中から数式領域を検出するステップと、

前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を用意しておき、前記数式領域内に含まれる前後の文字毎に正規化サイズとその中心位置の関係を算出し、その算出結果と、前記散布図とに基づいて、前記前後の文字間毎に、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造候補とその評価値から成るリンク候補を得る

文字間構造判定ステップと、

前記数式領域内に含まれる文字それぞれの文字高さの分布に基づいて予め決められた大域的評価条件を用意しておき、前記大域的評価条件と、前記リンク候補とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれかの文字間構造候補を選択しながら前記数式領域内の文字同士を矛盾なく接続するための最適な経路を探索し、その探索結果に基づいて前記各文字間の水平位置関係、下付添え字関係、上付添え字関係を認識するステップとを具備することを特徴とする文字認識方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、数式を含む文書イメージの認識に使用される数式認識装置および数式認識方法並びに文字認識装置および文字認識方法に関する。

【０００２】

【従来の技術】

数式を含む印刷文書の文字認識はあまり報告は多くはないものの、以前より少しずつ行われている。この分野の文字認識においては、文字は１次元的に並んでいるわけではなく、添え字やべき乗、分数線の上下など、２次元的な並びとなっている。従って、各文字の文字認識結果だけでなく、その文字が添え字、べき乗、分母・分子のどこにあるのかなど、数式における位置情報を判定する手段が必要となる。従って、この文字認識を計算機によって行う場合、その処理にかかる時間は数式を対象としない通常の文字認識と比べて多くなる。

【０００３】

これを実用的な時間で処理可能とした報告例に、以下に示す文献[1] [2]や文献[3]の手法がある。これらは、数式の中の文字の上下関係などをルールとして記述し、通常の文字、添え字、べき乗、分母・分子などの位置判定を行うことで、数式認識を実現している。

【０００４】

文献[1]　岡本正行,トワキョンドムサフィリハシム,「周辺分布特長を用いた

数式構造認識」,電子情報通信学会論文誌，J78-DII，No.2，pp.366-370(1995)

文献[2]　岡本正行,東裕之「記号レイアウトに注目した数式構造認識」,電子情報通信学会論文誌,J78-DII，No.3，pp.474-482(1995)

文献[3]　R. J. Fateman, T. Tokuyasu, B. P. Berman and N. Mitchell, "Optical Character Recognition and Parsing of Typeset Mathematics," Journal of Visual Communication and Image Representation, Vol 7,No. 1, pp.2-15 (1996)

【０００５】

【発明が解決しようとする課題】

しかしながら上記した従来技術においては、文字を局所的な特徴に基づいて通常の文字、添え字、べき乗、分母・分子などの位置判定を行っていたため、１箇所の位置判定が誤ると、その後の位置判定に大きく影響してしまうなどの問題があった。例えば、ある場所に存在する通常の文字を、誤って添え字と判定してしまうと、その誤判定された文字と同じ水平位置上に並んでいる以後の通常の文字についても、それらが全て添え字領域に存在する文字と誤って判定されてしまうなどの現象が発生する場合があった。つまり、局所的な誤認識が、数式の全体の構造認識を大きく崩してしまうのである。

【０００６】

また、上記した技術は、主に数式内部の文字認識に関するものであり、テキスト中に現れる数式を検出する方法については、単純に記号を検索するなどの仕組みに頼っていた。

【０００７】

そこで本発明は上記の問題を解決するためになされたものであり、数式を含む文書から高い精度で数式を認識することが可能な数式認識装置および数式認識方法並びに文字認識装置および文字認識方法を提供することを目的とする。

【０００８】

【課題を解決するための手段】

上述の課題を解決するため、本発明の数式認識装置は、数式を含む文書イメージの文字認識を行う文字認識手段と、正規表現により特定可能な単語種別毎にそ

れがテキストと数式に該当する可能性をそれぞれ示す評価値を定義した第１の知識辞書と、前記第１の知識辞書を参照して、前記文字認識手段によって得られた文字認識結果に含まれる各単語についてテキストおよび数式それぞれに該当する評価値を得る手段と、形式文法と前記各単語毎に算出されるテキストおよび数式それぞれの評価値とに基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路を探索し、その探索結果に基づいて数式に該当する単語を検出する数式検出手段とを具備することを特徴とする。

【０００９】

この数式認識装置においては、数式領域を通常の文字認識により認識すると、予期しない様々な文字が認識結果として出現することを考慮して、正規表現によって文字認識結果に含まれる様々な単語をその種別毎に分類し、且つその分類された単語種別毎に予めテキストと数式に該当する可能性をそれぞれ示す評価値を取得するための知識辞書が用意されている。この知識辞書を用いることにより、柔軟に各単語に対して評価値を与えることが可能となる。数式検出は、形式文法と、各単語毎に算出されるテキストおよび数式それぞれの評価値とに基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路を探索していくことによって行われる。これにより、数式領域を精度良く検出することができるので、数式を含む文書から高い精度で数式を認識することが可能となる。

【００１０】

また、本発明の数式認識装置は、数式を含む文書イメージの文字認識を行う文字認識手段と、前記文字認識手段によって得られた文字認識結果の中から数式領域を検出する手段と、前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を、前後の文字種類別に複数記憶する手段と、前記数式領域内に含まれる前後の文字毎に、正規化サイズとその中心位置の関係を算出し、その算出結果と、前記前後の文字の文字種類の関係に対応するサンプル情報とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係それぞれに該当する可能性を示すリンク候補を得る手段とを具備する

ことを特徴とする。

【００１１】

この数式認識装置においては、前後の文字種類別に異なる複数のサンプル情報が用意されており、水平位置関係、下付添え字関係、上付添え字関係を判定すべき文字間の文字種類に対応するサンプル情報を参照することにより、より高い精度で水平位置関係、下付添え字関係、上付添え字関係を判定することが可能となる。よって、数式内の文字の位置に関する判定誤り自体を大幅に低減することが可能となり、数式構造の認識効率を大幅に向上することができる。

【００１２】

また、本発明の数式認識装置は、数式を含む文書イメージの文字認識を行う文字認識手段と、前記文字認識手段によって得られた文字認識結果の中から数式領域を検出する手段と、前後の文字が水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の関係を示すサンプル情報を記憶する手段と、前記数式領域内に含まれる前後の文字毎に正規化サイズとその中心位置の関係を算出し、その算出結果と、前記サンプル情報とに基づいて、前記前後の文字間毎に、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造候補とその評価値から成るリンク候補を得る文字間構造判定手段と、前記数式領域内に含まれる文字それぞれの文字高さの分布に基づいて予め決められた大域的評価条件を記憶する手段と、前記大域的評価条件と、前記リンク候補とに基づいて、前記前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれかの文字間構造候補を選択しながら前記数式領域内の文字同士を矛盾なく接続するための最適な経路を探索し、その探索結果に基づいて前記各文字間の水平位置関係、下付添え字関係、上付添え字関係を認識する手段とを具備することを特徴とする。

【００１３】

このように、各文字間の局所的な関係の判定のみならず、大域的な評価条件を考慮して最終的に合計評価値が最大となるように数式領域内の文字同士を矛盾なく接続するための最適な経路が探索されるので、特定の文字間の位置判定にたとえ誤りが発生してとしても、それが数式全体の構造にまで影響を及ぼすことを防

止することが可能となる。

【００１４】

【発明の実施の形態】

以下、図面を参照して本発明の実施形態について説明する。

図１は本発明の一実施形態に係る文字認識システムの構成が示されている。この文字認識（ＯＣＲ）システム１１は、例えば科学技術文書などに代表されるような数式を含む印刷文書の認識を行うためのものであり、印刷文書をスキャナ装置１０を用いて読み取り、その文書内のテキスト領域および数式領域それぞれについての認識処理を行って、数式データとテキストデータとを含む電子化文書データを認識結果データ２０として出力する。読み取り対象の文書は印刷文書のみならず、既にイメージデータ化された数式混じりの文書イメージについても読み取り対象となる。

【００１５】

このＯＣＲシステム１１はコンピュータ上で実行されるソフトウェアとして実現されており、その機能モジュールとして、図示のように、レイアウト解析部１１１、通常文字認識部１１２、数式検出部１１３、数式認識部１１４、出力変換部１１５、数式・テキスト判定知識辞書２０１、品詞接続知識辞書２０２、文字サイズ散布図情報記憶部２０３、および大域的評価情報記憶部２０４を有している。これらの辞書および記憶部は半導体メモリや磁気ディスク等の記憶媒体に記憶されるものである。

【００１６】

認識処理は、１）文書イメージのスキャン、２）レイアウト解析処理、３）通常文字認識処理、４）数式検出処理、５）数式認識処理、６）出力変換処理、の順で行われる。本実施形態では、特に数式検出処理および数式認識処理の実現方法に特徴を有している。

【００１７】

数式検出処理および数式認識処理の具体的内容を詳述する前に、まず、処理の流れの概要について説明する。

【００１８】

先ず、数式を含む印刷文書をスキャナ装置１０で読み取ることにより、数式を含むページイメージが得られる。次いで、レイアウト解析部１１１によるレイアウト解析が行われ、ページイメージが図、表、文章領域に分割される。そして、文章領域に対して通常文字認識部１１２による通常文字認識処理が行われる。この通常文字認識処理では、ヒストグラムに基づく行の切り分けおよび文字の切り出し、そして１文字単位での文字認識が行われる。この後、文字認識結果に基づき数式検出部１１３による数式検出処理、および数式認識部１１４による数式認識処理が行われることになる。

【００１９】

数式検出部１１３による数式検出処理では、数式・テキスト判定知識辞書２０１、品詞接続知識辞書２０２が用いられる。数式・テキスト判定知識辞書２０１は、正規表現を用いて特定可能な単語の種別毎にそれがテキストと数式に該当する可能性をそれぞれ示す評価値を定義したものである。この数式・テキスト判定知識辞書２０１を参照することにより、単語毎にテキストと数式それぞれに関する評価値が求められる。

【００２０】

品詞接続知識辞書２０２は形式文法を規定したものであり、ここにはテキスト品詞と数式との間の接続関係の規則などが定義されている。この品詞接続知識辞書２０２で与えられる形式文法と、数式・テキスト判定知識辞書２０１の参照によって得られたテキストおよび数式それぞれに関する「評価値」とに基づいて、文字認識結果に含まれる単語間の最適な接続関係を判定することにより、文字認識結果が数式領域とテキスト領域へ分割される。

【００２１】

数式領域に含まれる文字・記号等は全て数式認識部１１４に送られ、そこで数式構造の認識処理が行われる。この数式構造認識処理では、数式をその構成要素に分解する処理がなされ、その後、各数式要素毎に、水平位置、下付添え字、上付添え字それぞれの関係などが調べられる。ここでは、文字サイズ散布図情報記憶部２０３内に記憶されている後述する複数種の文字サイズ散布図と、大域的評価情報記憶部２０４内に記憶されている後述する大域的評価条件とが用いられる

。サンプル情報である文字サイズ散布図は、前後の文字ペアが水平位置、下付添え字、上付添え字それぞれの関係にある場合におけるそれら前後の文字間における正規化サイズとその中心位置の分布の様子を示すものである。この文字サイズ散布図を参照することにより、数式要素内に含まれる文字間毎に、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造候補とその評価値から成るリンク候補が得られる。

【００２２】

大域的評価条件は、数式要素内に含まれる文字全てに関する大域的な評価に基づいて適切な文字間構造を決定するための条件式である。この大域的評価条件を用いることにより、各文字間の局所的な関係の判定のみならず、大域的な評価条件を考慮して、最終的に数式要素内の文字同士を矛盾なく関係付けするための最適な経路を探索する処理が行われる。

【００２３】

出力変換部１１５では、テキスト領域および数式領域それぞれについての認識結果等を合成して認識結果データ２０を出力する処理が行われる。

【００２４】

（数式検出方法）

以下、数式検出処理の具体的な方法について説明する。

本実施形態では、図２に示すように、以下の２つのステップ（Ａ１，Ａ２）からなる数式検出方法により、数式領域の検出を行う。この検出方法は、基本的に英文の文書からの数式検知を対象としている。

【００２５】

＜ステップＡ１：　数式／テキスト評価処理＞

このステップでは、通常の文字認識により得られた結果から、各単語を数式「Math」・テキスト「Text」として評価する。ここで「単語」とは認識結果のスペースで区切られた文字列をいう。図３は、この方法を示したものである。

【００２６】

図３の１行目は、実際に本システム１１へ入力された画像の例（Original Image）を示す。２行目はそれを通常文字認識部１１２により通常文字認識した結果

である（Rec gnized Result）。本実施形態の通常文字認識処理では、数式を認識する機能は実装されていないため、数式が現れるとその認識結果は予期しない様々な記号列として現れる。このステップＡ１では、この認識結果を入力として、各単語を数式「Math 」およびテキスト「Text 」としてそれぞれ評価する。認識結果の下の２行に「Math 」および「Text 」と示されている値は、こうして各単語を評価した結果の例を示す。本実施形態では、この処理を前述の数式・テキスト判定知識辞書２０１より検索することで行っている。図４に数式・テキスト判定知識辞書２０１のデータ例を示す。

【００２７】

図４において、番号１で示されている行は、「with 」という綴りの単語の品詞は前置詞(PP)で、「Math 」（数式）としての評価値が０、「Text 」（テキスト）としての評価値が１００であることを示す。同様に、番号２で示されている行は、「where 」という綴りの単語の品詞は代名詞(PN)で、「Math 」としての評価値が０、「Text 」としての評価値が１００であることを示す。番号３で示されている行は、「is 」という綴りの単語の品詞は動詞(V)で、「Math 」としての評価値が７０、「Text 」としての評価値が７０であることを示す。番号４で示されている行は、「a 」という綴りの単語の品詞は冠詞(ART)で、「Math 」としての評価値が９０、「Text 」としての評価値が９０であることを示す。このようにして、数式・テキスト判定知識辞書２０１には、科学技術文書などで通常使用されるほとんど全ての単語について、その綴り（文字コードの並び）、品詞、数式およびテキストそれぞれに関する評価値が予め登録されている。

【００２８】

さらに、本実施形態では、数式に対する認識結果は予期しない様々な記号列として現れることを考慮し、正規表現によって、様々な記号列に柔軟に対応できるようにしている。正規表現とは、単語の綴りをより柔軟に表現できるようにしたものであり、通常は検索システム等に使われている。この場合、正規表現における各記号は次の意味を表す。

【００２９】

. 任意の文字を示す

＊　直前の文字の０回以上の繰り返しを示す

（例　.＊ の場合、全ての文字列を示す）

[]　括弧内に指定された文字のいずれか１つを示す

（例　[a-z]の場合、a からz までのアルファベットの文字を示す）

^　次に指定した範囲以外の文字を示す

（例　[^ a-z]の場合、a からz 以外の文字を示す）

つまり、図４の番号５で示される行は、a からz 以外の文字、即ち何らかの記号を１文字含む単語であることを示す。この単語の品詞は名詞(N)で、「Math 」としての評価値が１００、「Text 」としての評価値が７０であることを示す。同様に番号６で示される行は、a からz 以外の何らかの記号を２文字含む単語であることを示しており、品詞は名詞(N)で、「Math 」としての評価値が１００、「Text 」としての評価値が４０である。番号７で示される行は、a からz 以外の何らかの記号を３文字含む単語であることを示しており、品詞は名詞(N)で、「Math 」としての評価値が１００、「Text 」としての評価値が２０である。番号８で示される行は、a からz までのアルファベット１文字を示しており、品詞は名詞(N)で、「Math 」としての評価値が９０、「Text 」としての評価値が４０である。なお、名詞(N)の品詞は該当する単語がテキストである場合を示している。

【００３０】

図４に示す数式・テキスト判定知識辞書２０１を行番号順に検索することにより、文字認識結果で得られた単語毎に品詞種別と、「Math 」および「Text 」それぞれについての評価値が得られる。

【００３１】

すなわち、図３に示されているように、単語［with ］については図４の番号１の知識により、「Math 」としての評価値が０、「Text 」としての評価値が１００として得られる。単語［ｆ ］については図４の番号８の知識により、「Math 」としての評価値が９０、「Text 」としての評価値が４０として得られる。単語［(,＼ ］の３文字については番号７の規則により、「Math 」としての評価値が１００、「Text 」としての評価値が２０として得られる。続く、単語であ

る［)=,＼ ］の４文字は「Math 」としての評価値が１００、「Text 」としての評価値が２０として評価していることを示す。ただし、図４にはこの例は示していない。同様に単語［where ］は図４の番号２の知識により、「Math 」としての評価値が０、「Text 」としての評価値が１００として得られる。単語［U ］は図４の番号８の知識により、「Math 」としての評価値が９０、「Text 」としての評価値が４０と評価される。同様に単語［is ］は図４の番号３の知識により、「Math 」としての評価値が７０、「Text 」としての評価値が７０として得られる。また、最後の単語［ａ ］は、図４の番号４の知識が番号８の知識よりも優先適用されるので、「Math 」としての評価値および「Text 」としての評価値が共に９０として得られる。

【００３２】

＜ステップＡ２：　最適パスの探索＞

次のステップＡ２では、評価した結果から最適パスを探索して接続する処理を行う。図５はこの様子を示したものである。このステップＡ２では、テキストのどの品詞がどの品詞に接続でき、またテキストのどの品詞が数式と接続できるかなどを示した前述の品詞接続知識辞書２０２を使用する。図６は品詞接続知識辞書２０２の実装例を示したものである。

【００３３】

図６において、１行目の「Text PP →Math」は、テキストの前置詞(PP)は後続する数式に接続できることを示している。また、２行目の「Math →Math」は、数式同士を接続できることを示している。３行目の「Math →Text PN」は、数式は後続するテキストの代名詞（PN）に接続できることを示している。４行目の「Text PN →Math」は、テキストの代名詞（PN）は後続する数式に接続できることを示している。５行目の「Text ART →Text N」は、テキストの冠詞（ART）は後続するテキストの名詞(N)に接続できることを示している。

【００３４】

品詞接続知識辞書２０２には接続可能な全ての組み合わせが登録されており、それ以外のものは接続できない。

【００３５】

最適経路の探索では、評価値を加算しながら各単語について、品詞接続知識辞書２０２の形式文法の規則に従って数式「Math 」／テキスト「Text 」のいずれかを選択しながら、可能な接続だけが辿られる。こうして、接続可能な全ての経路の中で、数式／テキストの評価値の合計が最も高くなる経路が探索される。簡単に言えば、例えば図５において単語[with]から次の単語[f]への接続可能な経路としては、単語[with]の「Math 」からは単語[f]の「Math 」と単語[f]の「Text 」とが存在し、また単語[with]の「Math 」からは単語[f]の「Math 」と単語[f]の「Text 」とが存在するが、選択経路の合計評価値が最も高くなるように、単語[with]の「Text 」から単語[f]の「Math 」への経路が選択されることになる。図５においては、最初の単語[with]から最後の単語[a]までの８単語を接続する際の最適経路として、「Text 」、「Math 」、「Math 」、「Math 」、「Text 」、「Math 」、「Text 」、「Text 」のルートが探索されたことが示されている。

【００３６】

この探索アルゴリズムは、ビームサーチ（または幅優先探索と言う）により実現できる。ビームサーチは動的計画法などで使用される良く知られたアルゴリズムであり、動的計画法において、最適経路としての可能性が低いと判断されたものを以後の処理から除外することで探索空間を圧縮し、計算量とメモリ量の低減を同時に実現できる効率化法である。

【００３７】

以上の探索処理の結果、各単語が数式「Math 」／テキスト「Text 」のいずれであるかが求まり、数式領域とテキスト領域とを検出することができる。図５では、

f

(,＼

)=,＼

U

の単語が数式「Math 」として判定され、それ以外の単語は全てテキスト「Text 」として判定されたことが分かる。数式「Math 」として判定された単語に対応

するイメージデータ内の領域が数式領域となり、またテキスト「Text 」として判定された単語に対応するイメージデータ内の領域がテキスト領域となる。

【００３８】

なお、本例では品詞を用いて接続をチェックするため、いわば正規文法で文法を記述しているのと等価であるが、実際には文脈自由文法など、より高度な形式文法で接続関係を記述することもできる。

【００３９】

従来のシステムでは、認識結果に括弧やイタリック体などの数式らしき記号が入っていればそれを数式と判定するなど、簡単なルールで判定しているものが多かった。従って、数式を認識した場合に認識結果として出現する様々な記号については対応できず、また例えば文書に［a ］という単語が存在した場合、それが冠詞であるか数式であるかを判定することも事実上不可能であった。本実施形態では、上述のように、各単語の評価値をチェックすることで、より正確に各単語が数式「Math 」であるかテキスト「Text 」であるかを判定できる。また形式文法をチェックしているので、例えば、冠詞であるテキスト［a ］に後続できるのはテキストの名詞のみであるという規則から、後ろに名詞が続かない［a ］については数式と判定することも可能となる。

【００４０】

（数式認識方法）

数式認識は、通常の文字認識と比べて、文字自体の認識の他に、添え字、べき乗、分母分子などの構造を調べる手法が必要となる。このうち本実施形態では文字自体の認識には、従来の文字認識と同一の方法を用いる。そして、数式構造を調べる方法については、図７に示すように、以下の４つのステップ（Ｂ１，Ｂ２，Ｂ３，Ｂ４）によって行われる。

【００４１】

＜ステップＢ１：　分母分子、左添え字、アクセント、根号、点類等の構造検出＞

このステップでは、数式領域のイメージデータから分数線や根号などを検出し、分母分子、根号内などをバラバラの式に分解する。同様に左添え字、アクセン

ト記号、点類などを検出し、それらを数式領域のイメージデータから消去する。

【００４２】

例えば、図８の様な数式が上記のようにして検出された数式領域に含まれている場合、点線で示すように４つの数式構成要素に分解され、且つ各数式構成要素毎に左添え字の削除（$^{3}a \rightarrow a$）、文字上の ^ , ¯ ,等のアクセント記号の削除（$xdx\hat{} \rightarrow xdx$）、さらに図８には示されていないが根号の削除（$\sqrt{a+b} \rightarrow a+b$）、点類の削除（$\dot{x} \rightarrow x$）などが行われる。

【００４３】

分母分子や左添え字、アクセント記号、根号、点類などの数式要素の判定は、上述の [1] [2] [3] などの文献でも比較的正確に行われており、多くの場合、局所的な位置関係に基づく判定式で判定可能である。そこで、これらの検出作業を単純な判定方法によりあらかじめ行っておくことで、以降のステップＢ２～Ｂ４の処理を、例えば下付添え字、上付添え字（べき乗）に関する処理に限定することができ、処理を高速化できる利点がある。

【００４４】

＜ステップＢ２：　文字認識＞

以降のステップＢ２～Ｂ４は、ステップＢ１により処理された、それ以上分数線やアクセント記号、左添え字、根号、点類などを含まない部分数式を対象に行う。

【００４５】

まず、ステップＢ２では、ステップＢ１によって得られた部分数式のイメージデータに対して黒連結成分の抽出がなされ、その各黒連結成分に対して文字認識が行われる。この結果、図９のような候補文字が得られる。図９は、$cx^{2}y^{3}$という部分数式のイメージデータを文字認識した場合の例であり、この文字認識により、各文字（黒連結成分）毎に大文字、小文字などが候補文字として得られる。

【００４６】

＜ステップＢ３：　リンク候補の生成＞

次のステップＢ３では、得られた候補文字の全てについて、図１０に示した関

係を用いて、各文字の接続可能性を調べる。

【００４７】

図１０は、前後の２つの文字間が水平位置関係、下付添え字関係、上付添え字関係のいずれに該当するかを判定するために用いる値（正規化サイズとその中心位置）を示したものである。図中、ｈ１，ｈ２で示した値は、それぞれ該当する文字の正規化高さ（正規化サイズ）である。正規化サイズとは、同一ライン上の文字についてはそれらが同じサイズ（高さ）を持つように大きさを補正したものである。

【００４８】

ここでは、アセンダー部分（例えば文字［ｄ］）とディセンダー部分（例えば文字［ｙ］）をあわせた文字全体の高さを正規化サイズとする。すなわち、ｈ１は、その文字の位置に「ｄ」と「ｙ」を重ねてタイプした場合の文字高さを示す。「ｄ」はアセンダー部分の上限にまで黒連結線分が延在している文字であり、「ｙ」はディセンダー部分の下限にまで黒連結線分が延在している文字である。例えば、図中に示した「ｘ」の場合、「ｄ」や「ｙ」と比べて背が低い。そこで、「ｘ」の実際の文字高さを一定倍することにより、「ｄ」と「ｙ」を重ね打ちした場合の正規化サイズｈ１を求めることができる。正規化サイズを求めるための倍率の値は、文字の種類毎に予め個々に規定されており、実際の文字サイズにその倍率を乗じることにより正規化サイズが求められる。例えば、小文字の「ｃ」についてはその上下方向に文字高さが広がるような倍率が用いられ、また大文字の「Ｃ」についてはその下方向にのみ文字高さが広がるような倍率が用いられることになる。

【００４９】

同様にして、添え字領域の文字「2」についてもその実際の文字サイズに対して、その文字「2」に対応する倍率を乗ずることにより、正規化サイズｈ２が求められる。通常、ベースライン上に存在する文字に比し、添え字領域に存在する文字の実サイズは小さいので、ベースライン上に存在する文字「ｘ」の正規化サイズｈ１よりも、添え字領域に存在する文字「2」の正規化サイズｈ２の方が小さくなる。

【００５０】

また、図１０において、ｃ１，ｃ２は、それぞれ正規化中心である。正規化中心とは同一ライン上の文字が同じ高さの中心位置を持つように中心位置を補正したものであり、ここでは、正規化した文字サイズを囲む外接矩形の中心ｙ座標を正規化中心とする。今、隣り合った文字の正規化高さと中心座標をそれぞれｈ１、ｃ１、ｈ２、ｃ２とすれば、

正規化サイズの関係　Ｈ＝（ｈ２／ｈ１）　×　1000

正規化中心の関係　　Ｄ＝｛（ｃ１－ｃ２）／ｈ１｝　×　1000

の関係をプロットすると、図１１の散布図が得られる。

【００５１】

図１１（Ａ）～（Ｄ）の４つの散布図（サンプル情報）は、水平位置にある文字のペアと、上付添え字の関係にある文字のペアと、下付添え字の関係にある文字のペアについて正規化サイズ・正規化中心の関係（Ｈ．Ｄ）を、前後の文字種類別に測定した結果を示している。図１１（Ａ）は連続する２つの文字が共にアルファベット類である場合の散布図である。ここで、アルファベット類とはアルファベット、ギリシャ文字、数字を示している。同様に、図１１（Ｂ）はアルファベット類と演算子とが前後する場合を示し、図１１（Ｃ）はインテグラルとアルファベット類とが前後する場合を示し、図１１（Ｄ）はΣ類とアルファベット類とが前後する場合を示している。

【００５２】

従って、ステップＢ２で調べた各候補文字間毎にＨ，Ｄを算出し、Ｈ，Ｄが、それらの文字種に対応する散布図上に示された多角形状領域のどれに属するかを評価することにより、水平位置関係、下付添え字関係、上付添え字関係の中で該当する可能性のある文字間構造候補とその評価値の組（ここではリンク候補と呼ぶ）を求めることが出来る。例えば前後の２文字間の正規化サイズ・正規化中心の関係（Ｈ．Ｄ）が図１１（Ａ）の多角形状領域Ｐ１，Ｐ２に含まれる場合にはそれらは上付添え字関係であると評価される（評価値はＰ２よりもＰ１に含まれる場合の方が高い）。また、多角形状領域Ｐ３，Ｐ４に含まれる場合にはそれらは下付添え字関係であると評価される（評価値はＰ４よりもＰ３に含まれる場合

の方が高い）。また多角形状領域Ｐ６，Ｐ５に含まれる場合にはそれらは水平位置関係であると評価される（評価値はＰ５よりもＰ６に含まれる場合の方が高い）。

【００５３】

図１２は生成されたリンク候補を分かりやすく示したものである。この図１２では、各リンク候補は、（親（左）候補文字、子（右）候補文字、接続の種類、評価値）を表している。なお、リンク候補は前後の２文字毎に行われるが、添え字領域が存在する文字を間に挟んでその前後にある２文字（図１２のｘ，ｙの関係）についてもリンク候補が生成される。

【００５４】

図１２に示すように、文字「ｃ」と文字「ｘ」とのリンク候補は、図１１（Ａ）の散布図を参照すると、

（ｃ，ｘ，水平，１００）

（ｃ，Ｘ，下，６０）

（Ｃ，Ｘ，水平，１００）となる。

【００５５】

この場合、（Ｃ，ｘ）の組は散布図からあり得ない。

【００５６】

また、文字「ｘ」と添え字文字「２」とのリンク候補は、図１１（Ａ）の散布図を参照すると、

（Ｘ，２，上，６０）

（ｘ，２，上，１００）

（ｘ，２，水平，２０）となる。

【００５７】

また、文字「ｘ」と添え字文字「２」を配慮した文字「ｙ」とのリンク候補は、図１１（Ａ）の散布図を参照すると、

（ｘ，ｙ，水平，１００）

（ｘ，Ｙ，下，６０）

（Ｘ，ｙ，水平，６０）

（２，ｙ，下，１０）

（２，Ｙ，下，５０）となる。

【００５８】

また、文字「ｙ」と添え字文字「３」とのリンク候補は、図１１（Ａ）の散布図を参照すると、

（ｙ，３，上，１００）

（Ｙ，３，上，５０）となる。

【００５９】

本実施形態では、図１１に示す散布図（サンプル情報）が、前後の文字種類別に４つある点が一つの特徴となっている。図１１に示した通り、各文字間の関係は前後の文字種類によって分布がかなり変化する。そこで本実施形態では、前後の文字の文字種類毎にこの図を用意して、判定対象の２文字の文字種類に対応した散布図を用いて添え字判定を行っている。

【００６０】

上述の文献[1] [2] [3]では、正規化された中心位置が親文字の中心当たりにあるか、上下にずれているかだけで、添え字判定を行っている。これは、図１１でいうと、縦座標だけを用いて添え字判定を行っていることになり、誤判定となる場合がかなりあることが分かる。これに対し、本発明では、大きさの比も組み合わせて２次元的な領域での散布図で判定を行い、更にそれを記号種毎の組み合わせで散布図を求めて判定を行っているため、添え字判定の精度が大幅に向上する。

【００６１】

次のステップを説明する前に、数式構造認識が何故最適経路問題になるかについて説明する。

【００６２】

即ち、数式の構造は木構造で表され、記号は１列に並ばないので、何故、最適「経路」を求める問題になるかは理解されていない。本発明では、ステップＢ３で作成したリンクネットワークから最適な数式構造を表す全域木を求めることにより達成される。「全域木を求めること」は「各文字の親文字への接続を定める

こと」になる。従って、

（親（左）候補文字、子（右）候補文字、接続の種類、評価値)の組を「リンク候補」と呼び、各文字矩形に、その文字を子とするリンク候補を全て持たせている。その上で、各文字矩形から１つずつリンク候補を選んでいけば１つの全域木が定まる。そのような選択は「経路」として見なすことが出来るので、最適経路問題になるという理屈になる。

【００６３】

＜ステップＢ４：　最適パスの探索＞

次いで、ステップＢ４では、ステップＢ３で文字間毎に生成されたリンク候補を、後ろから（又は前から）辿ることにより、それらリンク候補を接続する際の最適な経路が探索される。すなわち、各文字間毎の接続関係（水平位置関係、下付添え字関係、上付添え字関係）を考慮して、前後の文字間毎にいずれかのリンク候補を選択しながら文字同士を矛盾なく接続可能な経路の中で、最も合計評価値が最も高くなる経路が調べられる。この場合、各リンク候補で与えられる文字間毎の局所的な評価値のみならず、以下に示すように、該当する数式構成要素に含まれる文字それぞれの間の文字高さの分布等に基づく大域的な４つの大域的評価条件に基づいて、大域的評価値が最も高くなる経路が最適経路として決定される。

【００６４】

1．　経路内の各リンク候補の評価値の和を、大域的評価値とする。

【００６５】

2．　各文字の正規化サイズよりも、添え字領域にある文字の正規化サイズが大きければ大域的評価値を下げる。これは図１４（ａ）の場合に相当する。つまり、リンク候補によって添え字領域に存在すると判定された文字の正規化サイズが、他の文字それぞれの正規化サイズと等しいか、それよりも大きい場合には、大域的評価値を下げる。図１４（ａ）では、“ｂ”を添え字と同じ大きさと判断した場合で、“ｂ”の文字サイズが“ａ”と同じなので、大域的評価値を下げる。

【００６６】

3．　ベースライン上の文字と同じラインに近い文字が添え字領域にあれば、大

域的評価値を下げる。つまり、ベースライン上の文字と、図１１の散布図で狭領域（Ｐ２，Ｐ４，Ｐ６）に入る文字が添え字領域にあれば、大域的評価値を下げる。図１４（ｂ）では、“ｘ”を大文字の“Ｘ”と判断した場合で、ベースライン文字“Ａ”と同じラインに近い文字“Ｂ”が添え字領域にあり、大域的評価値を下げる。

【００６７】

４．ベースライン上のアルファベット類の正規化文字サイズが一定以上ばらついていれば、大域的評価値を下げる。これは図１４（ｃ）の場合に相当する。つまり、ベースライン上のアルファベット類が異なる正規化サイズを持つとき大域的評価値が下げられる。図１４（ｃ）は、“Ｃ”を小文字の“ｃ”に誤判定した場合で、その場合、“ｃ”の正規化サイズは“Ａ”の正規化サイズより大きくなり、大域的評価値を下げる。

【００６８】

このように、大域的評価条件とは、前後の文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれかのリンク候補を選択しながら数式内の文字同士を矛盾なく接続可能な経路における合計評価値を大域的な基準で修正し直すための条件である。大域評価値が最も高くなる最適な経路を探索するための探索アルゴリズムとしては、ビームサーチ（または幅優先探索と言う）を利用することができる。

【００６９】

図１３には、大域的評価値を考慮して決定された最適経路の一例が示されている。このようにして、各文字間毎に最適なリンク候補が選択され、各文字間毎に水平位置関係、下付添え字関係、上付添え字関係のいずれに該当するかが確定される。

【００７０】

上述の文献[1] [2] [3]の手法では、このような上記のような大域的評価値という考えが無かったため、一箇所でもベースライン上にある文字を添え字と間違うと、それ以降の文字が全て添え字になってしまう問題があった。これは、各文字の添え字・べき乗判定を、局所的な特徴のみに基づいて計算していることによる

ものであった。これに比し、本発明では経路を辿る時に大域的評価値を計算するため、１文字を誤って添え字と判定してしまったとしても、それ以後の文字を全て添え字としてしまうような現象が生じないという特徴を持つ。また、この大域的評価値計算方法を利用して、外部の装置により数式認識した結果を評価することもできる。これは複合判定などにも応用可能である。

【００７１】

そして、このようにして候補文字間の最適なつながりが決定された文字列に対してステップＢ１で削除した左添え字やアクセント記号、根号などを加えることにより、該当する数式構成要素に関する最終的な認識結果が得られる。ステップＢ２～Ｂ４の処理を数式構成要素毎に行うことにより、数式領域に関する最終的な認識結果が得られる。そして、テキスト領域の認識結果と数式領域に関する認識結果を合成することにより、数式を含む文章領域の認識結果データが得られる。

【００７２】

以上説明したように、本実施形態によれば、１）形式文法と各単語毎に算出されるテキストおよび数式それぞれの評価値とに基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路を探索することにより、数式領域を精度良く検出することが可能となる。２）前後の文字間における正規化サイズとその中心位置の関係を示す散布図を、前後の文字種類別に複数用意しておくことにより、高い精度で水平位置関係、下付添え字関係、上付添え字関係を判定することが可能となる。３）各文字間の局所的な関係の判定のみならず、大域的な評価条件を考慮して最適な経路が探索することにより、特定の文字間の位置判定にたとえ誤りが発生してとしても、それが数式全体の構造にまで影響を及ぼすことを防止することが可能となる。４）数式構成要素毎に分解して各数式構成要素から左添え字、アクセント記号、根号などを検出する処理を、リンク候補生成、最適パスの探索の前処理として事前に行うことにより、リンク候補生成の対象となる文字を減らすことができ、処理の効率化を図ること出来る。という効果が得られる。

【００７３】

なお、本実施形態のＯＣＲシステム１１の機能はすべてソフトウェアによって実現できるので、上述の各処理手順をコンピュータに実行させるプログラムを用意し、それをコンピュータ読み取り可能な記憶媒体に記憶すると共に、その記憶媒体を通じてコンピュータに導入して実行するだけで、本実施形態と同様の効果を容易に得ることができる。

【００７４】

また、本発明は、上記実施形態に限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で種々に変形することが可能である。更に、上記実施形態には種々の段階の発明が含まれており、開示される複数の構成要件における適宜な組み合わせにより種々の発明が抽出され得る。例えば、実施形態に示される全構成要件から幾つかの構成要件が削除されても、発明が解決しようとする課題の欄で述べた課題が解決でき、発明の効果の欄で述べられている効果が得られる場合には、この構成要件が削除された構成が発明として抽出され得る。

【００７５】

【発明の効果】

以上詳述した如く本発明によれば、数式を含む文書から高い精度で数式を認識することが可能となり、例えば科学技術文書の電子化等に有効に活用することができる。

【図面の簡単な説明】

【図１】

本発明の一実施形態に係るＯＣＲシステムの機能構成を示すブロック図。

【図２】

同実施形態における数式検出方法の手順を示すフローチャート。

【図３】

同実施形態の数式検出で行われる数式／テキスト評価処理を説明するための図。

【図４】

同実施形態で用いられる数式・テキスト判定知識辞書の例を説明するための図。

【図５】

同実施形態の数式検出で行われる最適パス探索処理を説明するための図。

【図６】

同実施形態で用いられる品詞接続知識辞書を説明するための図。

【図７】

同実施形態における数式認識方法の手順を示すフローチャート。

【図８】

同実施形態の数式認識で行われる数式分解の様子を示す図。

【図９】

同実施形態の数式認識で行われる候補文字の検出動作を説明するための図。

【図１０】

同実施形態の数式認識で行われる正規化サイズと正規化中心の算出処理を説明するための図。

【図１１】

同実施形態で用いられる散布図を説明するための図。

【図１２】

同実施形態において連続する文字間毎に生成されるリンク候補を説明するための図。

【図１３】

同実施形態の数式認識おける最適パス探索処理を説明するための図。

【図１４】

同実施形態の数式認識で用いられる大域的評価値計算のための条件を説明するための図。

【符号の説明】

１１…ＯＣＲシステム

１１１…レイアウト解析部

１１２…通常文字認識部

１１３…数式検出部

１１４…数式認識部

１１５…出力変換部

２０１…数式・テキスト判定知識辞書

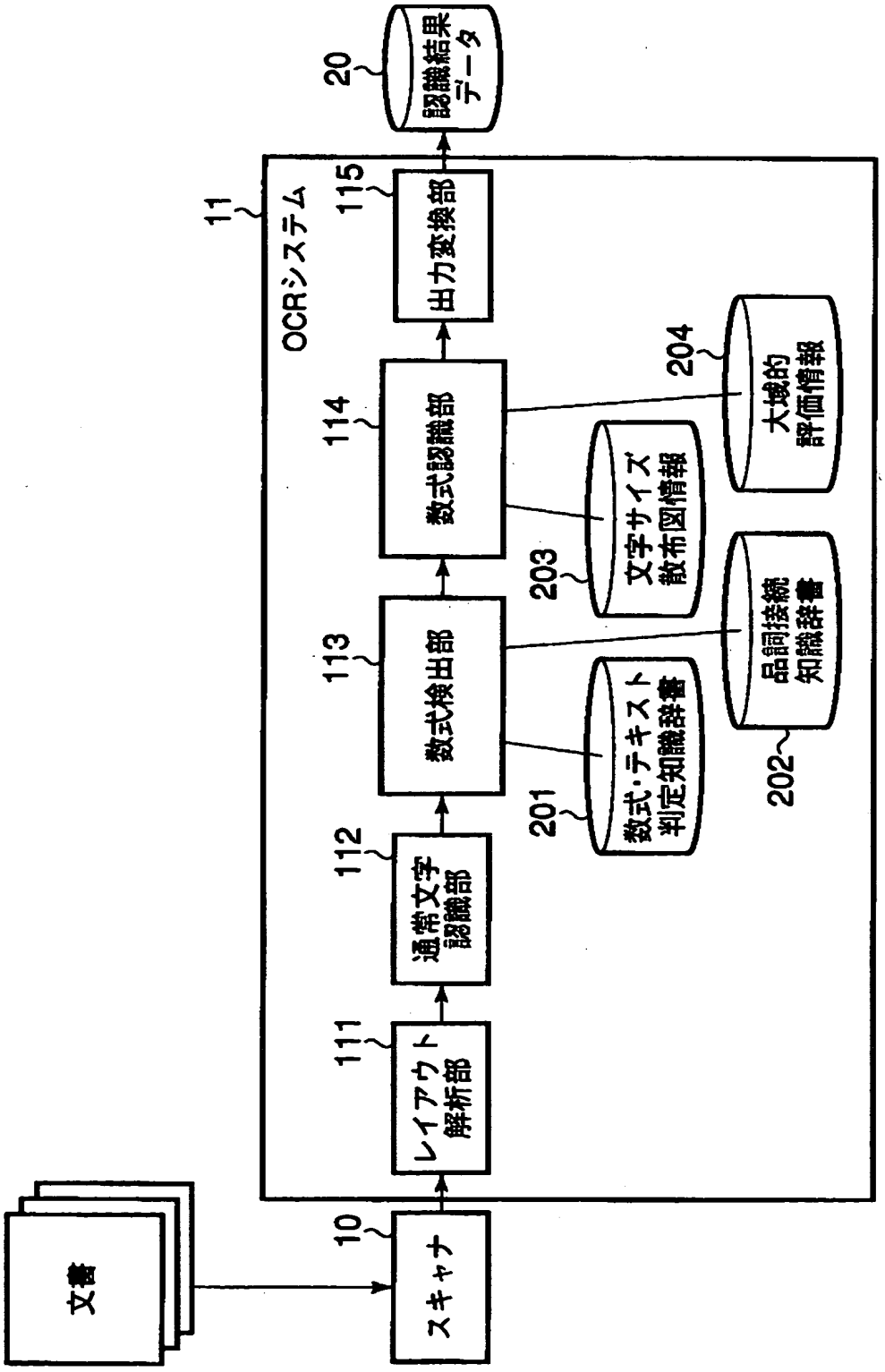
２０２…品詞接続知識辞書

２０３…文字サイズ散布図

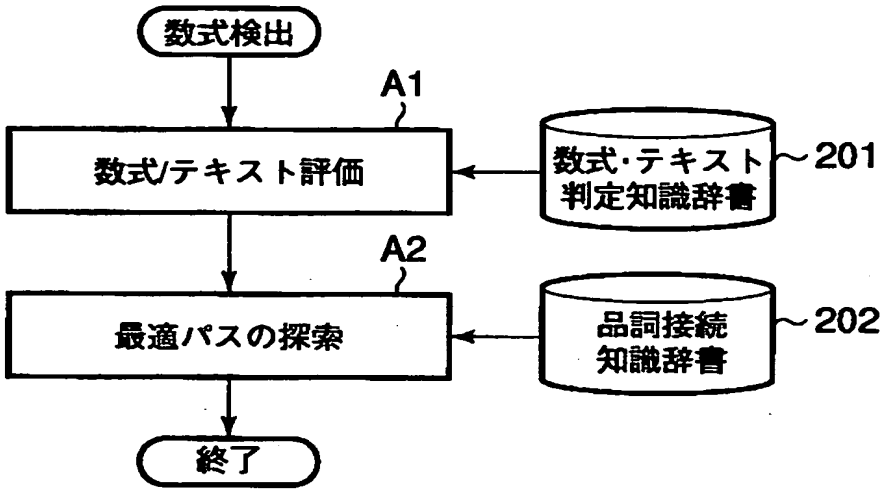
２０４…大域的評価情報

【書類名】　　図面

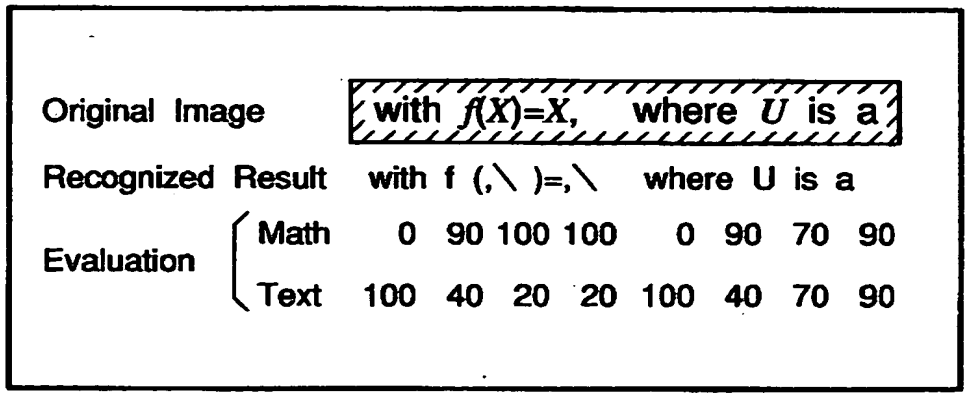
【図１】


文書
スキャナ
10
レイアウト解析部
111
通常文字認識部
112
数式検出部
113
数式認識部
114
出力変換部
115
OCRシステム
11
認識結果データ
20
数式・テキスト判定知識辞書
201
品詞接続知識辞書
202
文字サイズ散布図情報
203
大域的評価情報
204

【図２】


数式検出
A1
数式/テキスト評価
数式・テキスト
判定知識辞書
201
A2
最適パスの探索
品詞接続
知識辞書
202
終了

【図３】

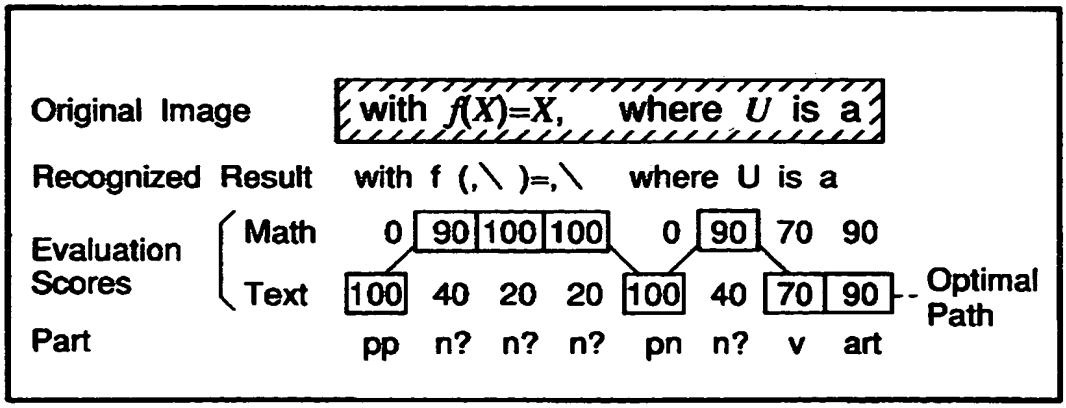

Original Image
with f(X)=X, where U is a
Recognized Result
with f (,\ )=,\ where U is a
Evaluation
Math 0 90 100 100 0 90 70 90
Text 100 40 20 20 100 40 70 90

【図４】

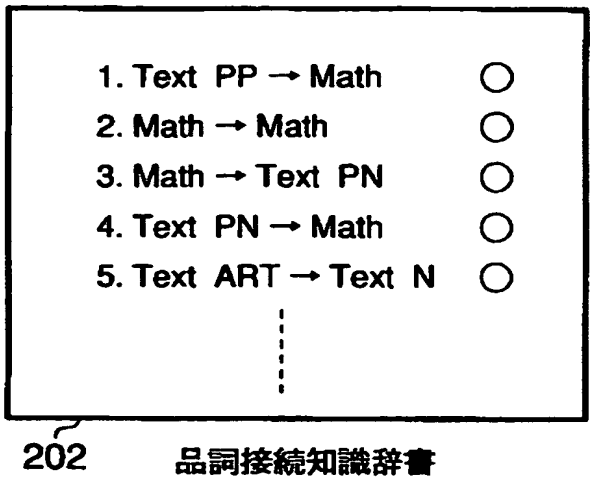
番号	単語	品詞	Math	Text
1.	with	PP	0	100
2.	where	PN	0	100
3.	is	V	70	70
4.	a	ART	90	90
5.	.*[^ a-z] .*	N	100	70
6.	.*[^ a-z] .*[^ a-z] .*	N	100	40
7.	.*[^ a-z] .*[^ a-z] .*[^ a-z] .*	N	100	20
8.	[a-z] *	N	90	40
9.	.*	N	100	5
⋮	⋮	⋮	⋮	⋮

201 数式・テキスト判定知識辞書

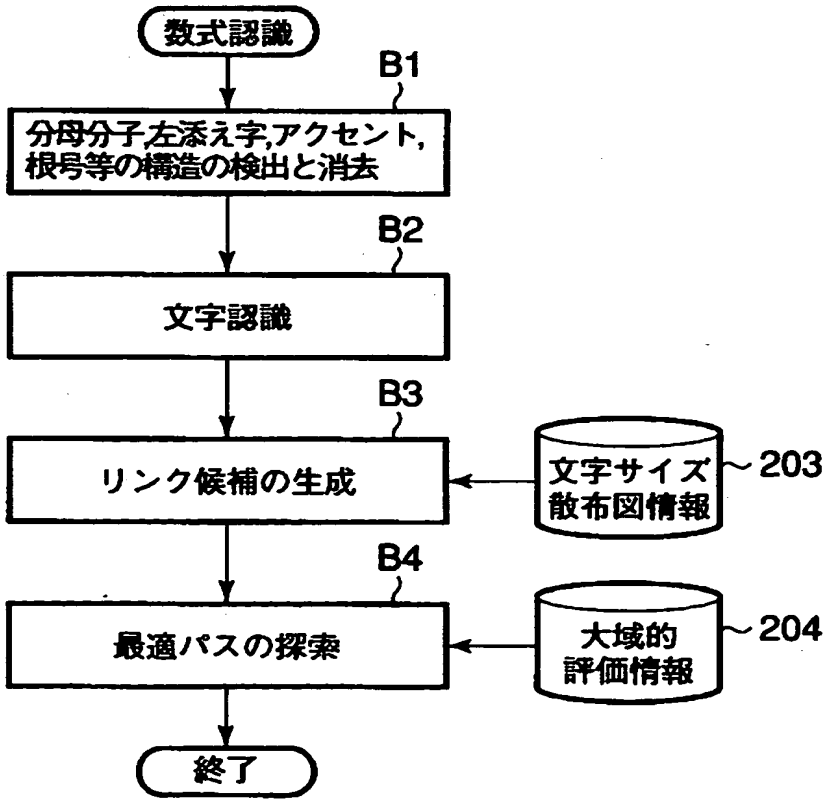
【図５】



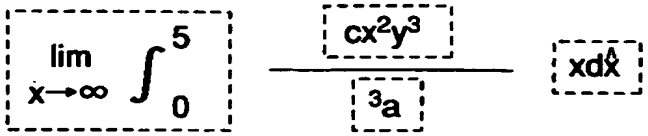
【図６】


1. Text PP → Math ○
2. Math → Math ○
3. Math → Text PN ○
4. Text PN → Math ○
5. Text ART → Text N ○
202
品詞接続知識辞書

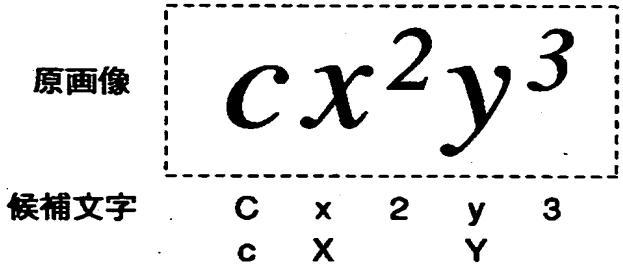
【図７】


数式認識
B1
分母分子,左添え字,アクセント,根号等の構造の検出と消去
B2
文字認識
B3
リンク候補の生成
文字サイズ散布図情報
203
B4
最適パスの探索
大域的評価情報
204
終了

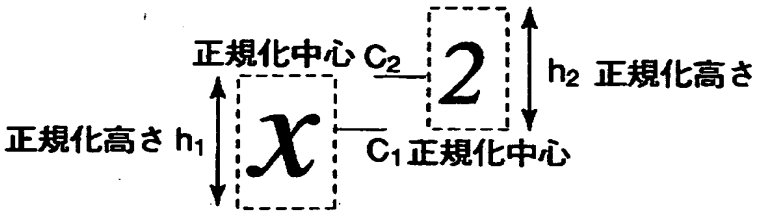
【図８】


lim
x→∞
5
0
cx²y³
³a
xdx

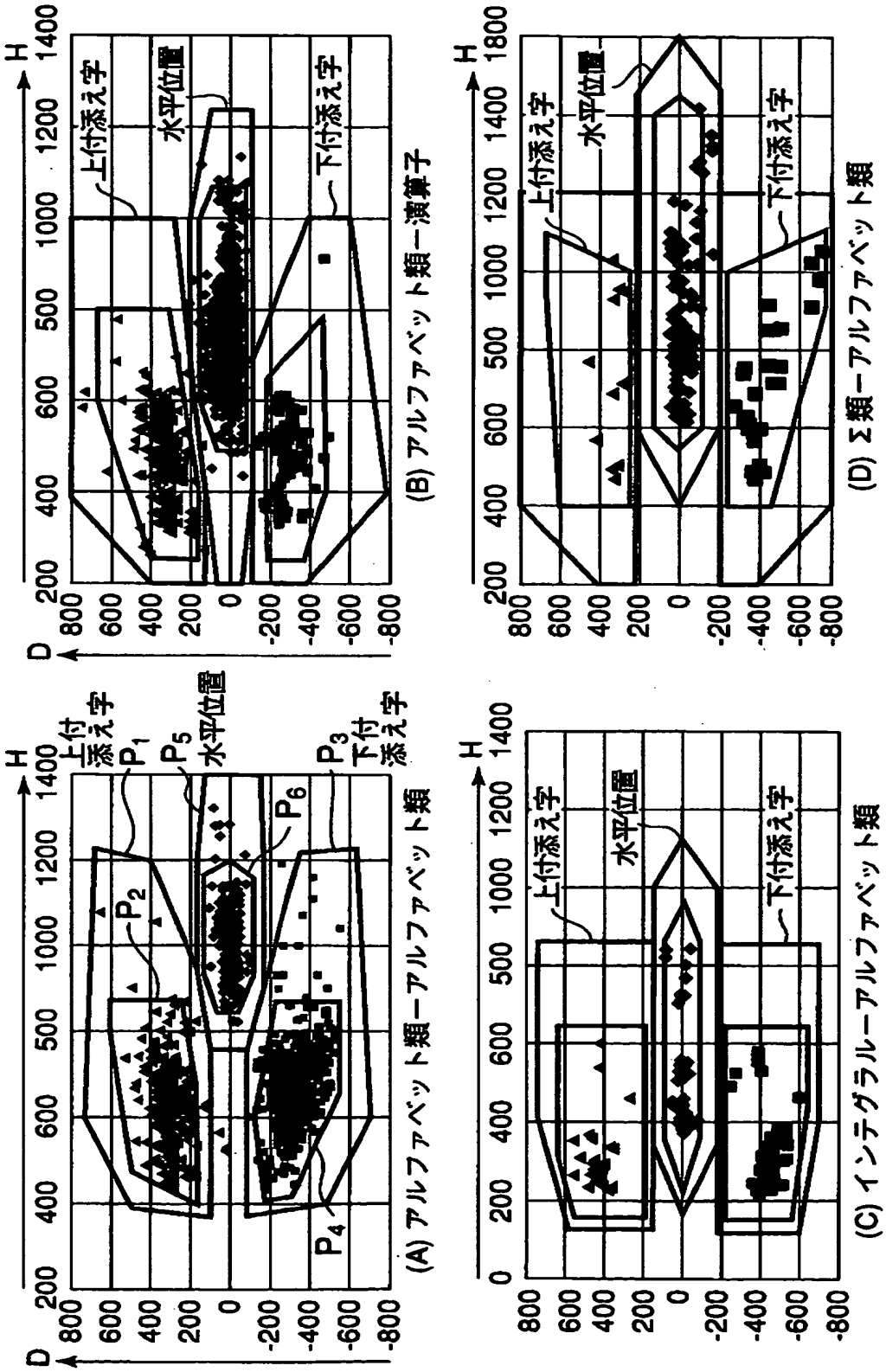
【図９】


原画像
cx²y³
候補文字
C x 2 y 3
c X Y

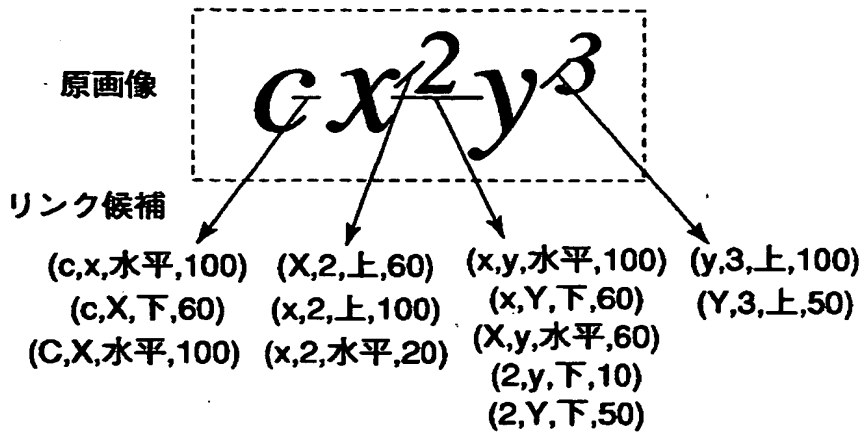
【図１０】


正規化中心 C2
2
h2 正規化高さ
正規化高さ h1
x
C1 正規化中心

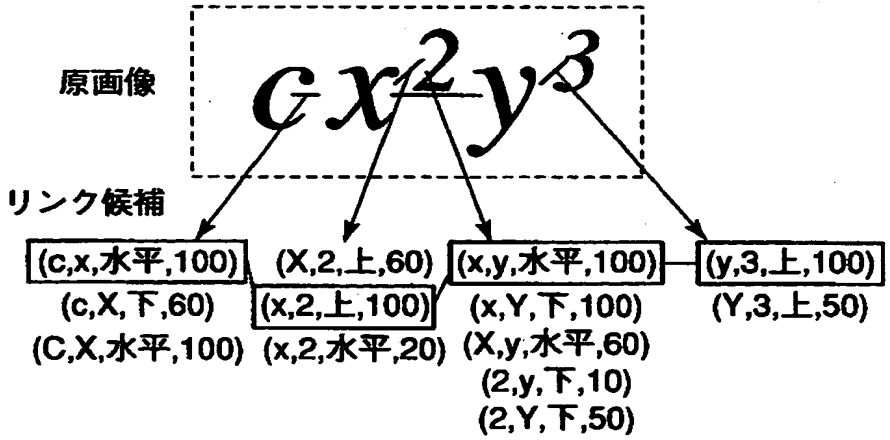
【図１１】


H
D
上付添え字
水平位置
下付添え字
(B) アルファベット類－演算子
(D) Σ類－アルファベット類
(A) アルファベット類－アルファベット類
(C) インテグラル－アルファベット類
P1
P2
P3
P4
P5
P6

【図１２】


原画像
cx^2y^3
リンク候補
(c,x,水平,100)
(c,X,下,60)
(C,X,水平,100)
(X,2,上,60)
(x,2,上,100)
(x,2,水平,20)
(x,y,水平,100)
(x,Y,下,60)
(X,y,水平,60)
(2,y,下,10)
(2,Y,下,50)
(y,3,上,100)
(Y,3,上,50)

【図１３】


原画像
cx^2y^3
リンク候補
(c,x,水平,100)
(c,X,下,60)
(C,X,水平,100)
(X,2,上,60)
(x,2,上,100)
(x,2,水平,20)
(x,y,水平,100)
(x,Y,下,100)
(X,y,水平,60)
(2,y,下,10)
(2,Y,下,50)
(y,3,上,100)
(Y,3,上,50)

【図１４】

(a) $a^2+b \rightarrow a^{2+b}$

"b"を添え字と同じ大きさと判定した場合で、"b"の文字サイズが"a"と同じなので大域的評価値を下げる

(b) $A_x+B_y \rightarrow A_{X+By}$

"x"を大文字の"X"と判断した場合でベースライン文字"A"と同一ラインに近い文字"B"が添え字領域にあり、大域的評価値を下げる

(c) $A_x+C_x \rightarrow A_{x+cx}$

"C"を小文字の"c"と誤判定した場合で、その場合"C"の正規化サイズは"A"の正規化サイズより大きくなり、大域的評価値ちを下げる

【書類名】　　要約書

【要約】

【課題】数式を含む文書から高い精度で数式を認識することが可能なＯＣＲシステムを実現する。

【解決手段】数式検出部１１３では、形式文法と各単語毎に算出されるテキストおよび数式それぞれの評価値とに基づいて、単語毎にテキストおよび数式のいずれかを選択しながら単語間を接続するための最適な経路が探索され、数式領域が検出される。続く数式認識部１１４では、前後の文字種類別に異なる複数の散布図を用いることにより、水平位置関係、下付添え字関係、上付添え字関係についての判定がなされる。そして、各文字間の局所的な関係の判定のみならず、大域的な評価条件を考慮して最適な経路を探索することにより、文字間毎に生成されたリンク候補の中から最適な経路が決定され、文字間の添え字関係が確定される。

【選択図】　　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００３０７８〕

１．変更年月日　　１９９０年　８月２２日
［変更理由］　新規登録
住　所　神奈川県川崎市幸区堀川町７２番地
氏　名　株式会社東芝

２．変更年月日　　２００１年　７月　２日
［変更理由］　住所変更
住　所　東京都港区芝浦一丁目１番１号
氏　名　株式会社東芝

出　願　人　履　歴　情　報

識別番号　　〔５０１０９２１４０〕

１．変更年月日　　２００１年　３月　７日
［変更理由］　新規登録
住　所　　福岡県福岡市東区箱崎６丁目１０番１号　九州大学内
氏　名　　鈴木　昌和